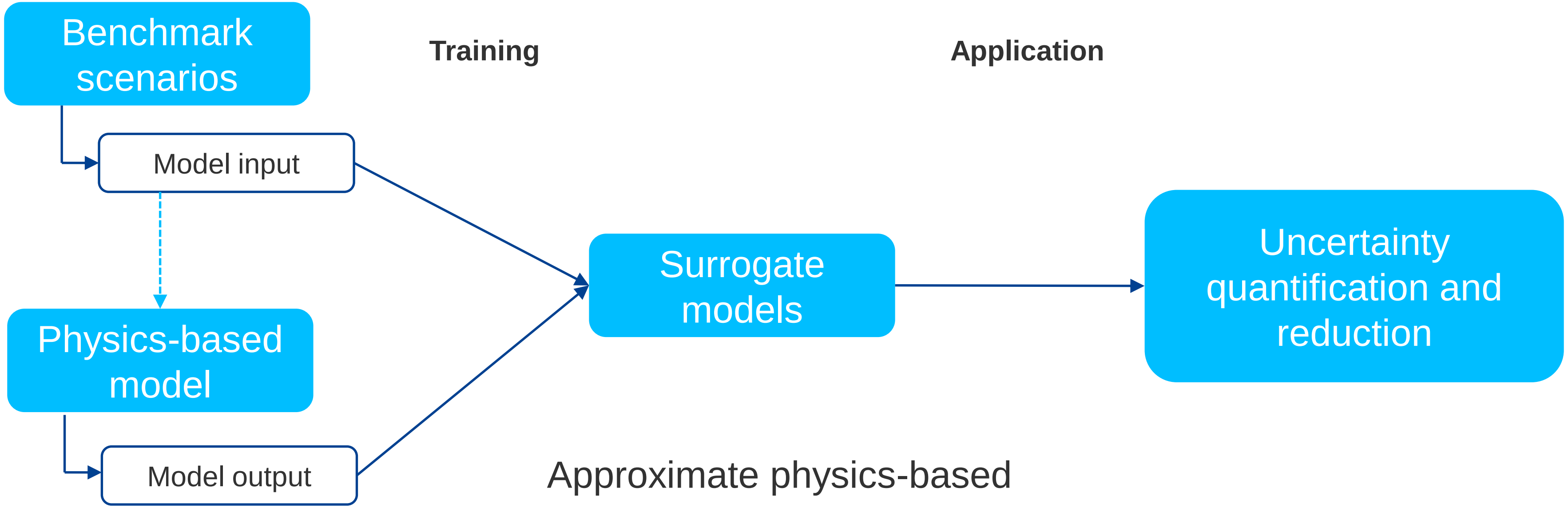
University of Stuttgart
Germany

Surrogate model generation using Bayesian active learning

Maria Fernanda Morales Oreamuno, M.Sc.

Recap

Surrogate modelling in the context of Smart Monitoring



Approximate physics-based model

Run in a fraction of the time

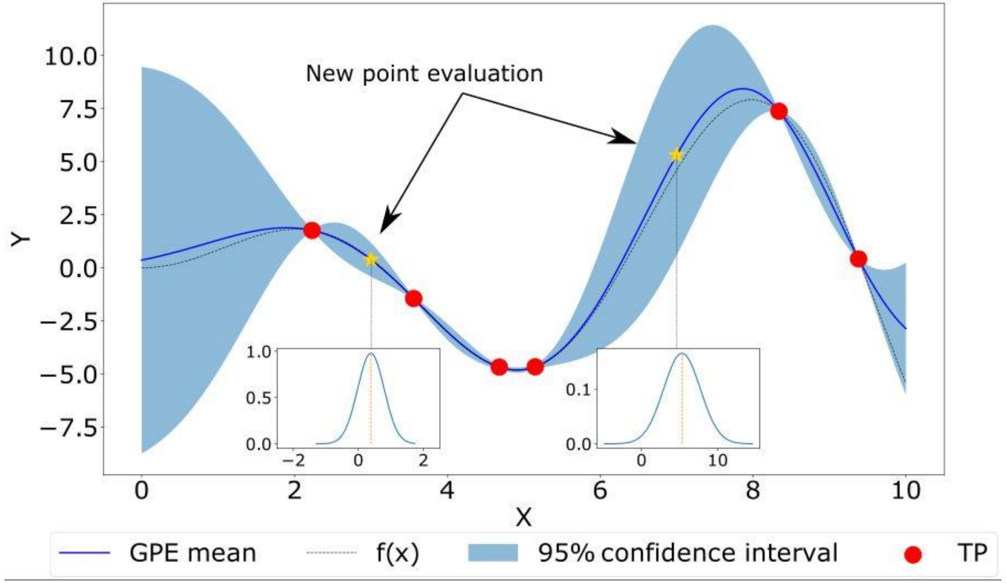
Surrogate modelling: Gaussian process regression

Recap

- Approximates the full-complexity model (simulator)
- Trained through input-output pairs (TP), generated by the simulator
- Predictions for any (future) parameter combinations are described by:
  - Mean
  - Variance

\+ Reduces computation time

\- Induces approximation error



1D input – 1D output example using Gaussian process regression

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

Crevillen-Garcia, D., Wilkinson, R. D., Shah, A. A., & Power, H. (2017). Gaussian process modelling for uncertainty quantification in convectively-enhanced dissolution processes in porous media. *Advances in water resources*, *99*, 1-14.

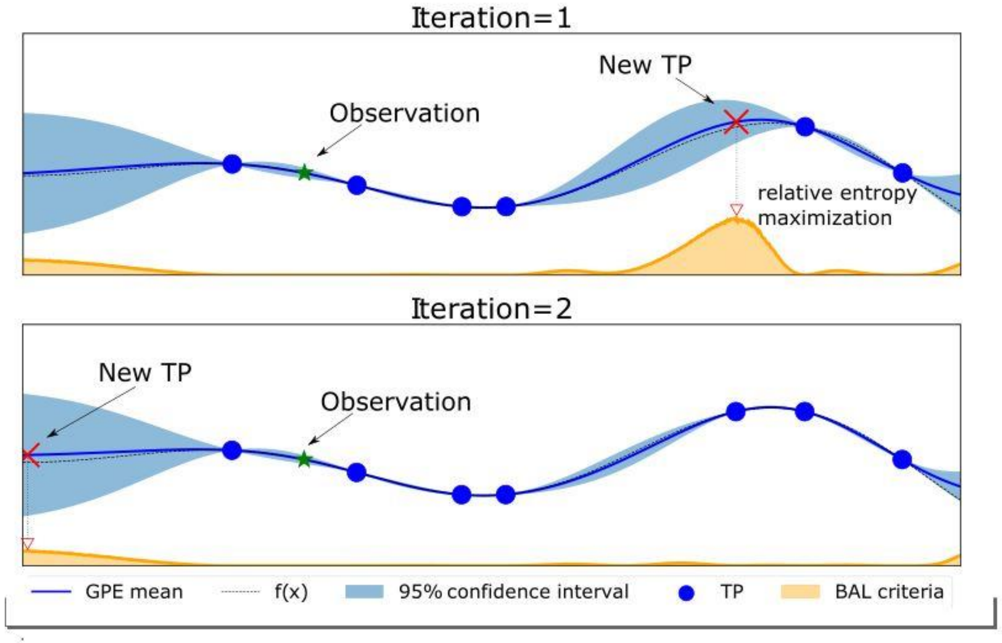
Surrogate modelling: Bayesian Active Learning

Recap

- Methodology to select training points
- Uses field observations
- Criteria:
  - **Bayesian inference** → obtain information from measurements

    +

  - **Information theory:** uncertainty associated with predictions



Information theory scores as training point selection criteria using Bayesian active learning

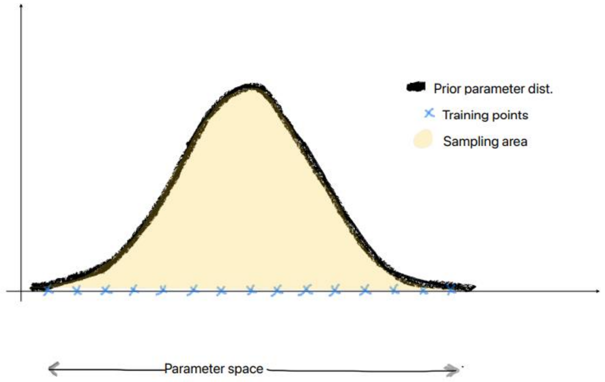
Oladyshkin, S., Mohammadi, F., Kroeker, I., & Nowak, W. (2020). Bayesian[3] active learning for the gaussian process emulator using information theory. *Entropy*, *22*(8), 890.

Zhao, H., & Kowalski, J. (2022). Bayesian active learning for parameter calibration of landslide run-out models. *Landslides*, 1-13.

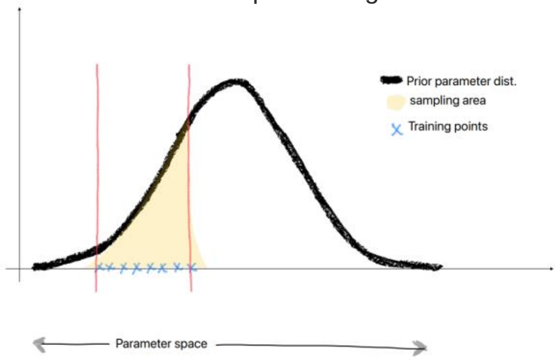
Surrogate modelling: Bayesian Active Learning

Recap

- Methodology to select training points
- Uses field observations
- Goals:
  - **Improve** the surrogate in a region, where it is more likely that the true parameters are
  - **Reduce** the number of training points needed



Space-filling



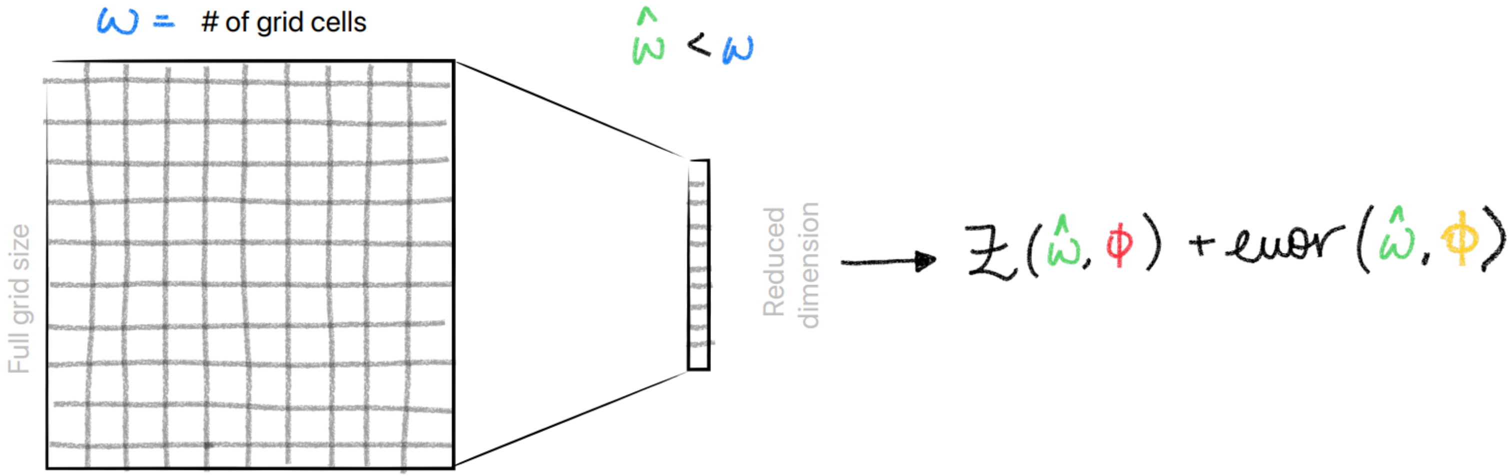
Goal of BAL

Oladyshkin, S., Mohammadi, F., Kroeker, I., & Nowak, W. (2020). Bayesian[3] active learning for the gaussian process emulator using information theory. *Entropy*, *22*(8), 890.

Zhao, H., & Kowalski, J. (2022). Bayesian active learning for parameter calibration of landslide run-out models. *Landslides*, 1-13.

Initial tests

GPR + BAL

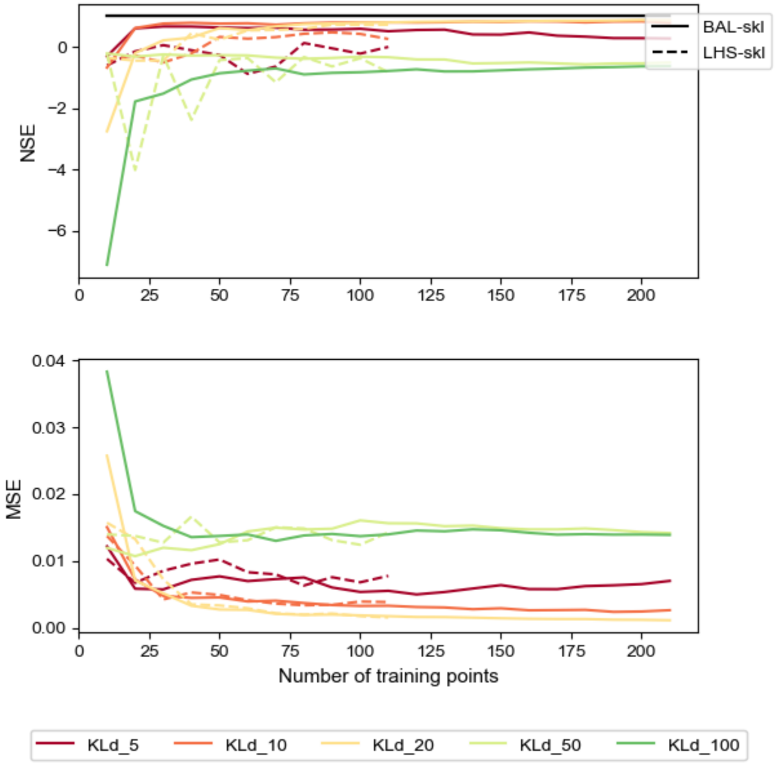


Initial goals:

- Heterogeneity and input dimension reduction → **effectiveness?**
- Bayesian active learning to select training points
- Evaluation criteria → **Bayesian and non-stochastic methods**

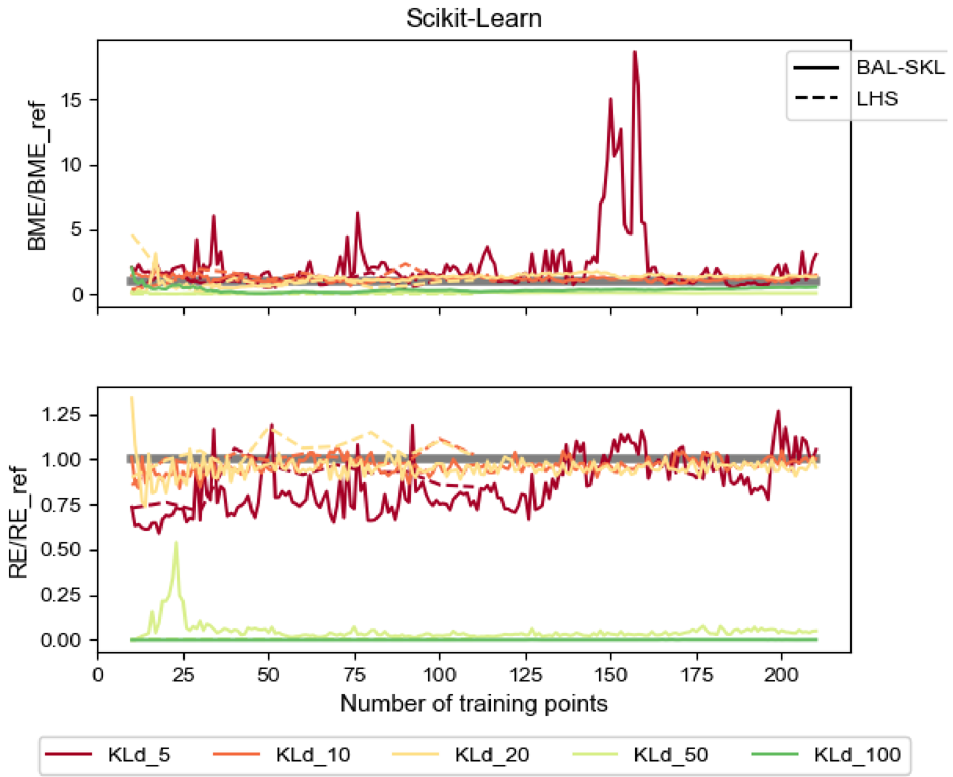
GPR + BAL Tests

Evaluation criteria for different input dimensions



Cross-validation evaluation criteria

Compared against simulator runs



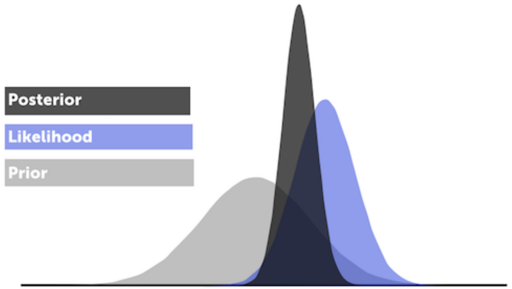
Bayesian evaluation criteria

Compared against observation data

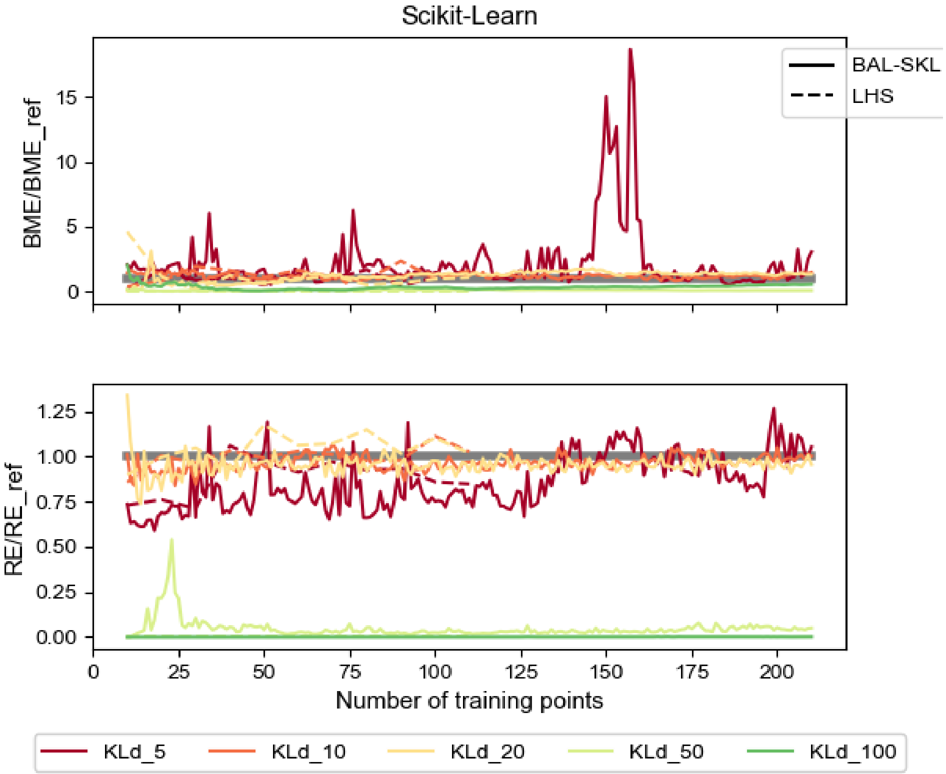
GPR + BAL Tests

Issues encountered

- Problems with Bayesian criteria for large number of observations
  - BAL with high observation space due to likelihood function –> posterior sampling methods



Outlook: posterior sampling methods to strengthen Bayesian criteria



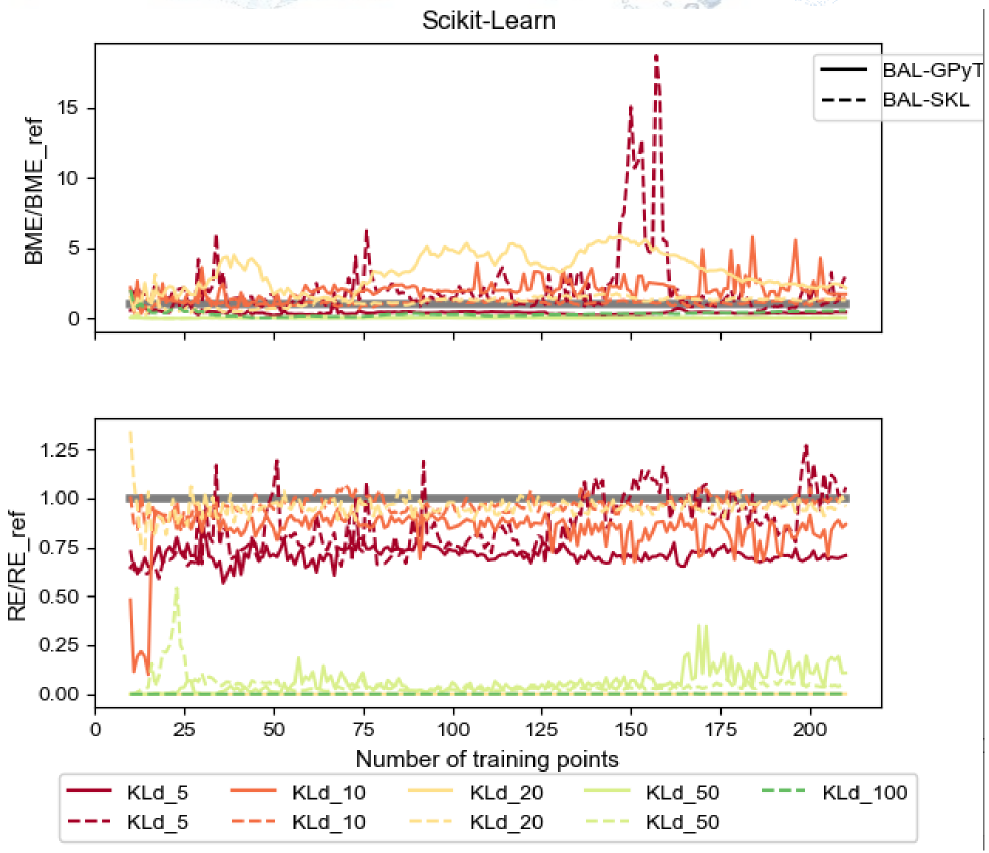
Bayesian evaluation criteria

Compared against observation data

GPR + BAL Tests

Issues encountered

- GPR libraries: Scikit-Learn and GPyTorch
  - Scikit-Learn shows a better overall fit for smaller input dimensions
- High input dimensionality: problems with GPR
  - **Outlook**: look into (arbitrary) Polynomial Chaos Expansion for surrogate modelling
    - Sensitivity analysis
    - Combine it with GPR (error quantification)



Bayesian evaluation criteria

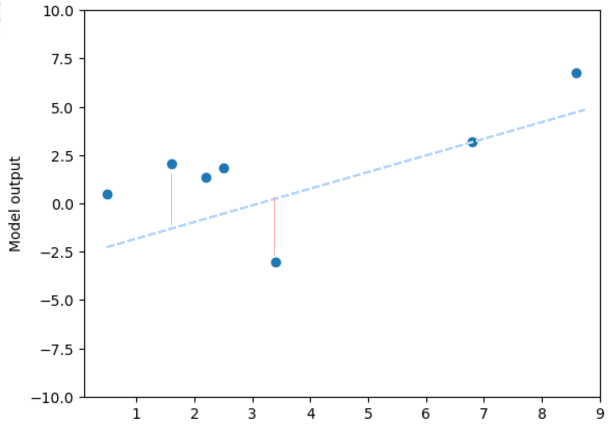
Comparing 2 different GPR Python libraries

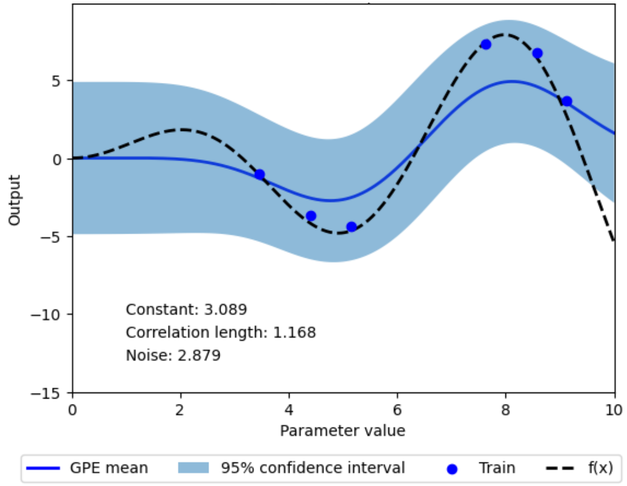
GPR + BAL Tests

Issues encountered

- GPR libraries: Scikit-Learn and GPyTorch
  - Scikit-Learn shows a better overall fit for smaller input dimensions

- High input dimensionality: problems with GPR
  - **Outlook**: look into (arbitrary) Polynomial Chaos Expansion for surrogate modelling
    - Sensitivity analysis
    - Combine it with GPR (error quantification)





1D example using GPR with noise

Surrogate modelling

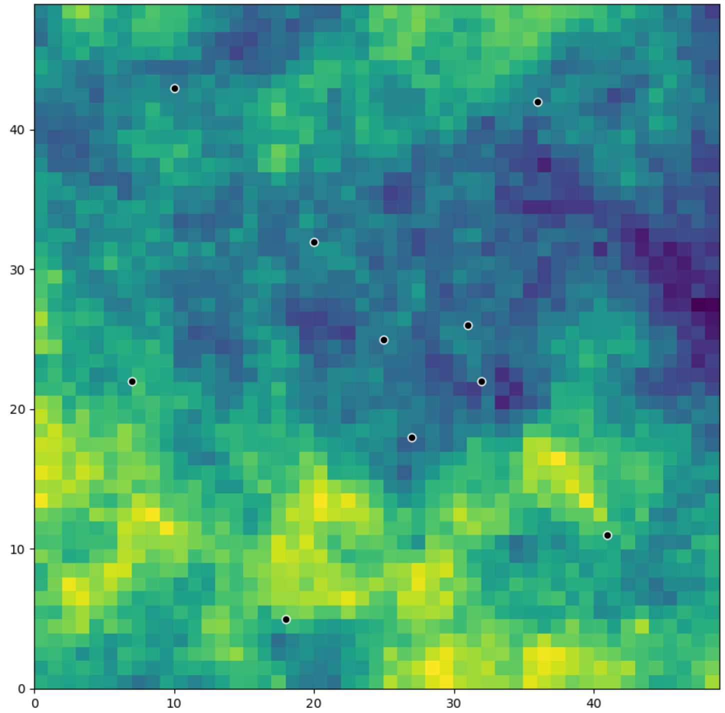
Output dimension reduction

- High output dimension
  - Train GPE for each cell and/or each time step

  High computational time

- **Outlook:** Output dimension reduction
  - Principal component analysis (PCA)
  - Kriging
  - Neural networks

- Useful for optimal design of experiments



Configuration of 2D groundwater model, with a 50 m x 50 m grid. Allows for different input-output scenarios

Outlook

- With heterogeneity + input dimension reduction
  - Quantify the error induced by using a smaller input data set

- With Bayesian Active Learning
  - Look into strengthening selection criteria (MCMC sampling, likelihood-free methods?

- Test arbitrary Polynomial Chaos Expansion (aPCE) for surrogate generation
  - Do a sensitivity analysis for input parameter space
  - Couple with GPR for uncertainty quantification

- **Implement methods on a simple application-specific case (including relevant processes)**

University of Stuttgart
Germany

Thank you for your attention!

Maria Fernanda Morales Oreamuno

maria.morales@iws.uni-stuttgart.de

University of Stuttgart
Stochastic Simulation and Safety Research for Hydrosystems (LS^3)
iws-ls3.uni-stuttgart.de

References

Crevillen-Garcia, D., Wilkinson, R. D., Shah, A. A., & Power, H. (2017). Gaussian process modelling for uncertainty quantification in convectively-enhanced dissolution processes in porous media. *Advances in water resources*, *99*, 1-14.

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., and Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. In Advances in Neural Information Processing Systems. https://gpytorch.ai/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Van-derplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830. https://scikit-learn.org/stable/modules/gaussian_process.html

References

Oladyshkin, S., Mohammadi, F., Kroeker, I., & Nowak, W. (2020). Bayesian^3 active learning for the gaussian process emulator using information theory. *Entropy*, *22*(8), 890.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

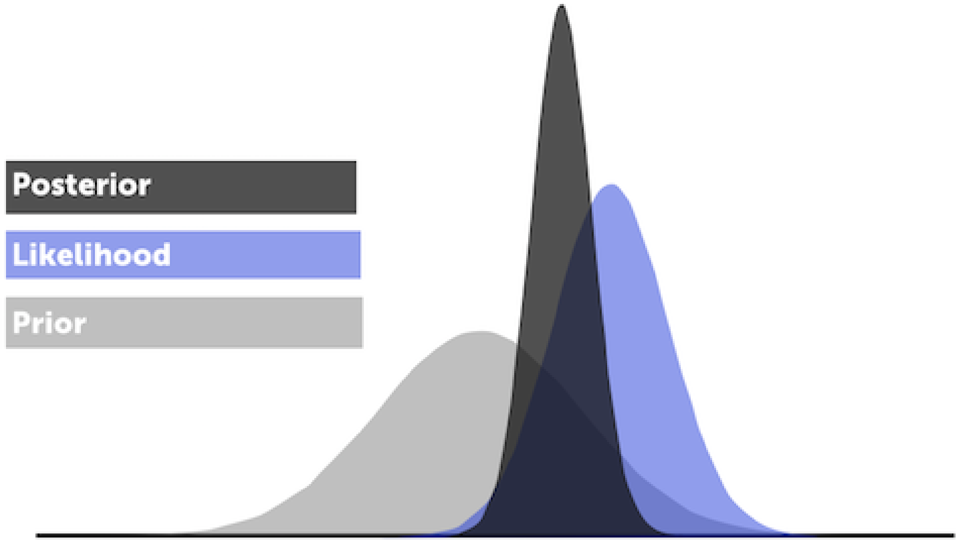
Zhao, H., & Kowalski, J. (2022). Bayesian active learning for parameter calibration of landslide run-out models. *Landslides*, 1-13.

Bayesian inference

- Bayes' theorem: update a prior state of knowledge to a posterior based on observation data

posterior likelihood prior

$$p(\omega|y_o) = \frac{p(y_o|\omega)p(\omega)}{p(y_o)}$$

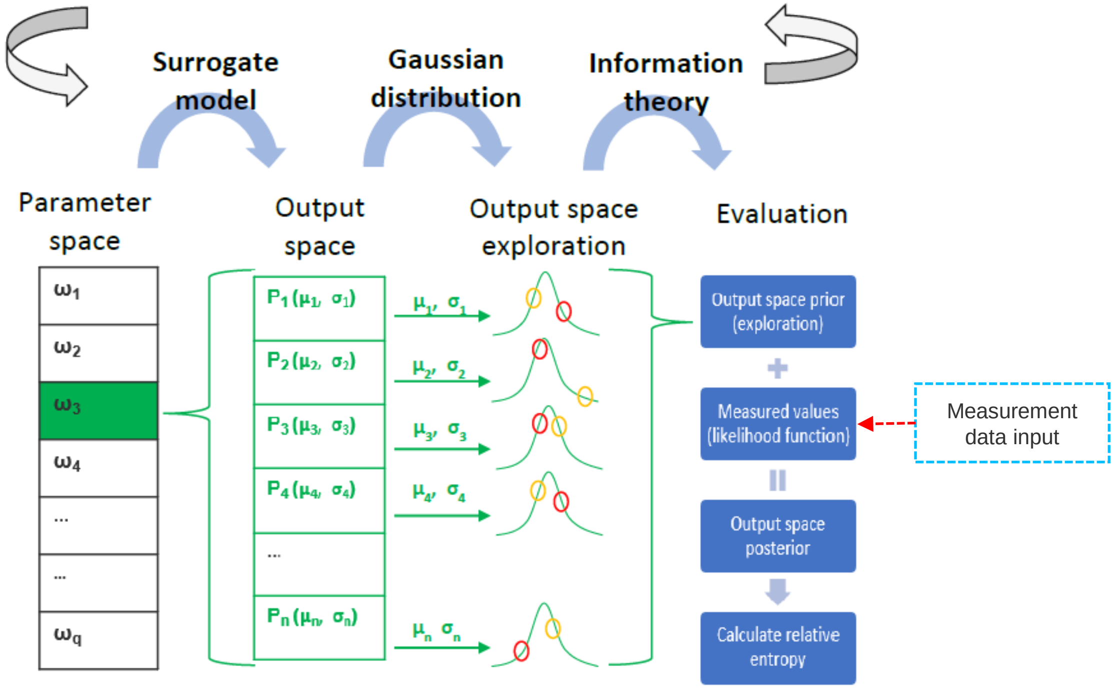


Bayesian inference: updating a prior to a posterior using observation data, through a likelihood function

Source: Oladyshkin (2022), IWS lecture, University of Stuttgart

Bayesian active learning

- (Bayesian) Active Learning allows to select training points located in regions of **high posterior likelihood**:
  - to improve the surrogate model prediction
  - reduce the number of total training points needed.
- For each iteration of the surrogate training, one selects the parameter set $\boldsymbol{\omega_i}$ which presents the highest gain in information as the next training point



Source: Acuna Espinoza (2021)